American Journal of Botany

RESEARCH ARTICLE

BRIEF COMMUNICATION

A novel chloroplast gene reported for flagellate plants

Michael Song[1,5,*], Li-Yaung Kuo[2,3,*], Layne Huiet[4], Kathleen M. Pryer[4], Carl J. Rothfels[1], and Fay-Wei Li[2,3]

Manuscript received 6 October 2017; revision accepted 22 November 2017.

[1] University Herbarium and Department of Integrative Biology, University of California, Berkeley, California 94720, USA

[2] Boyce Thompson Institute, Ithaca, New York 14853, USA

[3] Section of Plant Biology, Cornell University, Ithaca, New York 14853, USA

[4] Department of Biology, Duke University, Durham, North Carolina 27708, USA

[5] Author for correspondence: (e-mail: michael_song@berkeley.edu)

*These authors contributed equally.

Citation: Song, M., L.-Y. Kuo, L. Huiet, K. M. Pryer, C. J. Rothfels, and F.-W. Li. 2018. A novel chloroplast gene reported for flagellate plants. *American Journal of Botany* 0(0): 1–5.

doi: 10.1002/ajb2.1010

PREMISE OF THE STUDY: Gene space in plant plastid genomes is well characterized and annotated, yet we discovered an unrecognized open reading frame (ORF) in the fern lineage that is conserved across flagellate plants.

METHODS: We initially detected a putative uncharacterized ORF by the existence of a highly conserved region between *rps16* and *matK* in a series of *matK* alignments of leptosporangiate ferns. We mined available plastid genomes for this ORF, which we now refer to as *ycf94*, to infer evolutionary selection pressures and assist in functional prediction. To further examine the transcription of *ycf94*, we assembled the plastid genome and sequenced the transcriptome of the leptosporangiate fern *Adiantum shastense* Huiet & A.R. Sm.

KEY RESULTS: The *ycf94* predicted protein has a distinct transmembrane domain but with no sequence homology to other proteins with known function. The nonsynonymous/synonymous substitution rate ratio of *ycf94* is on par with other fern plastid protein-encoding genes, and additional homologs can be found in a few lycophyte, moss, hornwort, and liverwort plastid genomes. Homologs of *ycf94* were not found in seed plants. In addition, we report a high level of RNA editing for *ycf94* transcripts—a hallmark of protein-coding genes in fern plastomes.

CONCLUSIONS: The degree of sequence conservation, together with the presence of a distinct transmembrane domain and RNA-editing sites, suggests that *ycf94* is a protein-coding gene of functional significance in ferns and, potentially, bryophytes and lycophytes. However, the origin and exact function of this gene require further investigation.

KEY WORDS *Adiantum shastense*; open reading frame (ORF); plastome; RNA editing; *ycf94*

Since plastid genomes (plastomes) were first sequenced from tobacco (Shinozaki et al., 1986) and *Marchantia* (Ohyama et al., 1986), hundreds of complete land-plant plastomes have become available. This plethora of plastome data has not only facilitated molecular phylogenetics (reviewed in Soltis et al., 2012), but has also enabled comparative studies that examine patterns of plastome rearrangements and molecular evolution (reviewed in Jansen and Ruhlman, 2012). It is now widely accepted that the plastome gene content among land plants is generally similar, with approximately 80 protein-coding genes, 30 tRNAs, and four rRNAs (Jansen and Ruhlman, 2012).

Although there are occasional gene losses or transfers to the nuclear genome, dramatic deviations from this basic content are rare and are mostly observed in parasitic and mycoheterotrophic plants (reviewed in Graham et al., 2017). The function of plastome genes has also been well characterized in land plants, with only six loci still labeled as *ycfs* (hypothetical chloroplast open reading frames; Jansen and Ruhlman, 2012). Some of these *ycf*s have already been found to be involved in photosystem assemblies (*ycf3* and *ycf4*; Boudreau et al., 1997) or protein import machinery (Kikuchi et al., 2013).

There are several bioinformatics approaches to assess whether newly identified open reading frames (ORFs) encode for a functional protein, such as identifying sequence conservation across species, selection pressure, and functional domains (Hsu and Benfey, 2017). RNA-editing sites are also a marker of potential gene function (Kugita et al., 2003). RNA editing changes the nucleotide sequence

of primary transcripts so that it differs from what was originally transcribed from the DNA. Typical RNA editing in plants involves C-to-U conversions, although U-to-C conversions are also common in certain plants (Schallenberg-Rüdinger and Knoop, 2016). Taken together, these approaches provide substantial evidence for translated ORFs.

Although land-plant plastomes are well annotated, we recently encountered an unknown ORF across ferns that we designate here as *ycf94* according to the plastome gene-naming convention (Hallick and Bairoch, 1994). We find evidence to support that this ORF encodes a transmembrane protein and that it is conserved among ferns, lycophytes, and bryophytes (but is not present in seed plants), possibly a conservation of an ancient ORF in non-seed plants.

MATERIALS AND METHODS

Plastome assembly

To obtain the complete plastome of *Adiantum shastense* (voucher: Huiet 201 UC2030515), DNA was extracted from fresh leaf tissue using a DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) and sequenced on one lane of Illumina HiSeq 4000 (500 base pair [bp] library insert and 150 bp paired-end reads). Raw reads were processed by Scythe (Buffalo, 2011) to remove adaptor sequences and by Sickle (Joshi and Fass, 2011) to trim low-quality bases. We then used NOVOPlasty (Dierckxsens et al., 2016) to assemble the plastome, which was validated by read-mapping in Geneious version 7.1.8 (Kearse et al., 2012). Genome annotation was done in Geneious using the *A. capillus-veneris* plastome (Wolf, 2003) as the reference, followed by manual adjustments. The raw reads were deposited in NCBI SRA (SRP123721), and the plastome GenBank accession number is MG432483.

Sequence analyses

To compare the ratio of nonsynonymous (K_a) to synonymous (K_s) nucleotide substitution rates of *ycf94* with other plastome protein-coding genes, we used the dataset from Li et al. (2016), which includes 48 gene alignments from 21 fern plastomes. We excluded *Dipteris conjugata* from the dataset because its plastome sequence is incomplete and the fragment containing *ycf94* is missing. K_a/K_s for each locus was calculated by codeml in the PAML package (Yang, 2007) with F3X4 codon frequencies, no clock, and a single K_a/K_s value (omega parameter) across the tree. To compare *ycf94* sequence divergence to those of other plastid coding genes, we analyzed *ycf94*, *matK*, and *rbcL* from the fern genus *Deparia*, which is the only genus to date with a comprehensive species-level *ycf94* sampling (Kuo et al., 2016). The sequence alignments were obtained from the *rps16-matK* intergenic spacer (IGS) + *matK* + *rbcL* + *trnL-L-F* alignment used in Kuo et al. (2016; available at the Dryad Digital Repository: https://doi.org/10.5061/dryad.450dp); *ycf94* spans position 998 to 1215 in that alignment. In total, 48 *Deparia* taxa were included. For each of the three codon positions, uncorrected pairwise sequence divergence was calculated using MEGA 7 (Tamura et al., 2013), with "p-distance," "d: Transitions + Transversions," "Uniform rates," and "Pairwise deletion" selected. We identified putative *ycf94* homologs in other plant lineages using Blastn (Altschul et al., 1990). The TMHMM server (Krogh et al., 2001) was used to predict a transmembrane motif if present.

Identification of RNA-editing sites

RNA was extracted from young leaves of the same *A. shastense* individual using the Sigma Spectrum Plant RNA kit (Sigma-Aldrich, Darmstadt, Germany), which has been shown to perform well for ferns (Rothfels et al., 2013). To capture the organelle transcripts, we used an Illumina Ribozero leaf kit to prepare the library, which was sequenced on Illumina HiSeq 2000 (150 bp paired-end reads). Raw reads were processed as described above, and the cleaned reads were mapped to the *A. shastense* plastome using Tophat 2 (Kim et al., 2013). We then used ChloroSeq (Castandet et al., 2016) and custom scripts to identify RNA-editing sites and to calculate RNA-editing efficiencies (defined as the percentage of total mapped RNA reads that are edited).

RESULTS

The complete *A. shastense* plastome contains a circular sequence of 150,414 bp, which comprises an 82,113 bp large single-repeat region, a 21,539 bp small single-repeat region, and two 23,381 bp inverted repeat regions. The gene content and gene order in the *A. shastense* plastome are the same as that of *A. capillus-veneris* (Wolf et al. 2003).

In *A. shastense*, *ycf94* is a 219 bp ORF located in the plastome between *rps16* and *matK*; this position appears to be conserved across Polypodiales, Cyatheales, Salviniales, and Schizaeales (Fig. 1), where *matK* has lost its flanking *trnK* exons (Kuo et al., 2011). This ORF was likewise found in all fern plastomes analyzed, where it varies in size from 150 bp to 270 bp. Homologs of *ycf94* were not found in seed plants. In lycophytes, homologs were found in all Lycopodiales (*sensu* PPG I, 2016) for which complete plastome sequences were available, but no homologs were found in Isoetales or Selaginellales (Appendix S1; see Supplemental Data with this article). Among nonvascular plants, homologs of *ycf94* were found in each of the three main lineages (six of eight available moss plastomes; three of six liverwort plastomes; both hornwort plastomes; Appendix S1); *ycf94* appears to be homologous with ORF51 in hornworts and ORF50 in liverworts, which were both reported but not characterized hypothetical proteins (Shimada and Sugiura, 1991; Kugita et al., 2003).

TMHMM (Krogh et al., 2001) analyses indicate that fern *ycf94* homologs have a distinct transmembrane domain (Appendix S2). Transmembrane domains were likewise identified as highly probable in the bryophyte homologs, with a similar single transmembrane domain predicted; for lycophytes, there is more variability in this region, and two transmembrane domains were predicted in *Huperzia* (Appendix S2). We then investigated whether *ycf94* is a target of RNA editing in the *A. shastense* plastome and found that most of the RNA-editing sites are in the gene regions (Appendix S3) and that there are four RNA-editing sites in *ycf94* (Table 1). Similarly, in the hornwort *Anthoceros angustus* Stephani, three RNA-editing sites were found in the homologous coding region (Kugita et al., 2003).

Plastome-wide K_a/K_s values across 20 fern species indicate that *ycf94* is under purifying selection similar to other protein-coding genes. Among *Deparia* species, the first and second positions of *ycf94* were more conserved than the third position (Fig. 2), similar to the patterns observed in *rbcL* and *matK*. The results from all analyses are consistent with our hypothesis that *ycf94* is a protein-coding gene in fern, bryophyte, and lycophyte plastomes.

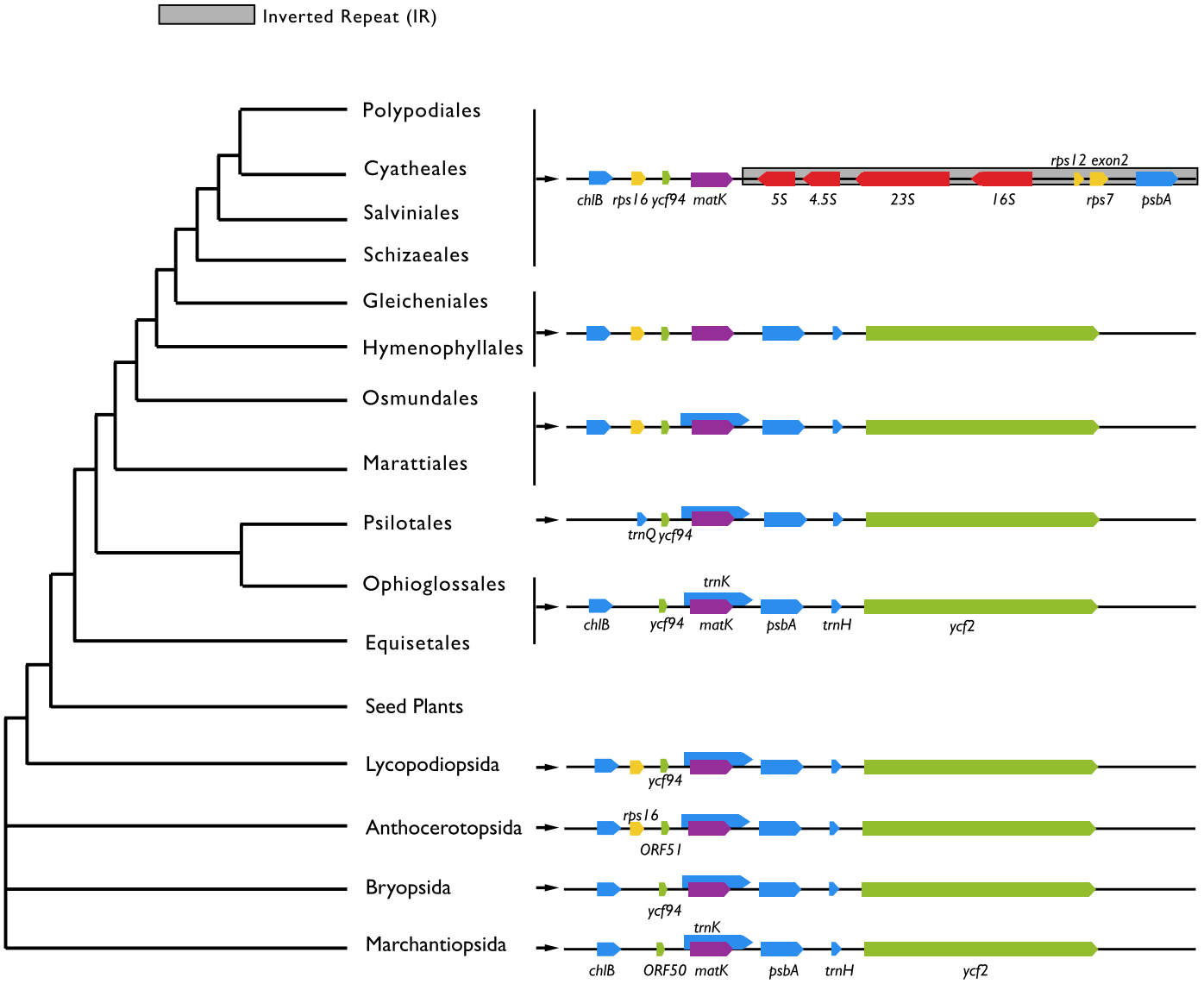

FIGURE 1. Map of the general location of *ycf94* (and its homologs ORF50 and ORF51 in hornworts and liverworts) in the plastome of flagellate plants—*ycf94* appears to be absent from the plastomes of seed plants and heterosporous lycophytes. Arrowheads indicate the direction of transcription; *trnK* is shown above *matK* to illustrate the *matK* ORF in the group II intron of *trnK* (Duffy et al., 2009; Kuo et al., 2011); rRNA genes are shown in red, and ycfs are shown in green. The cladogram on the left is based on PPG I (2016) for the relationship of ferns and on Wickett et al. (2014) for the uncertainty in the relationship among bryophytes. Note that some genes are omitted for clarity.

TABLE 1. RNA-editing sites in *ycf94* in *Adiantum shastense*.

Position	Editing direction	Amino acid change	Editing efficiency
2	C → T	T → M	0.10
50	C → T	S → F	0.35
56	C → T	P → L	0.24
75	C → T	Synonymous	0.45

DISCUSSION

Despite the sequence variability both upstream and downstream of *ycf94* across the taxa examined (Fig. 1), we found the *ycf94* sequences to be conserved throughout the available fern plastid sequences (Appendix S4) as well as in many species of mosses, hornworts, liverworts, and homosporous lycophytes examined—strongly suggesting the origin of *ycf94* in a common ancestor of extant land plants. Sequence conservation across many taxa, as well as the finding that *ycf94* has a consistently predicted transmembrane domain, provides evidence that this ORF encodes a potentially functional protein. Another line of evidence is the presence of RNA-editing sites in the genic regions. These results (Appendix S3) are consistent with the expectation that RNA editing is characteristic of protein-coding genes (Gray and Covello, 1993), and, importantly, we found four C-to-U RNA-editing sites in *ycf94* (Table 1)—three of which result in nonsynonymous substitutions (Appendix S5). When we compared the K_a/K_s ratios of *ycf94* across fern plastomes, we found them to be comparable to those of other plastome protein-coding genes (Fig. 2). Likewise, when we examined variation across codon

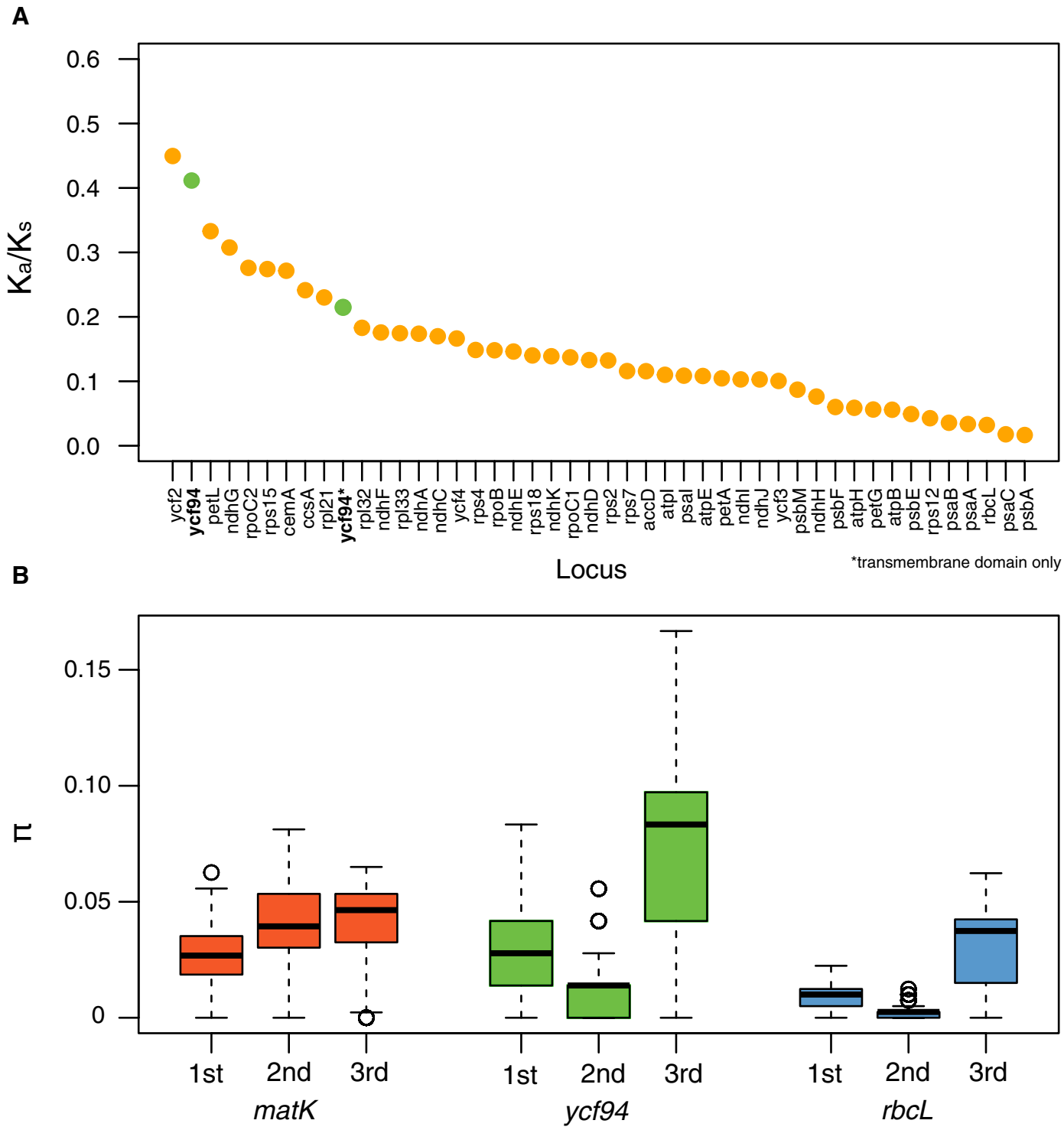

FIGURE 2. Comparison of *ycf94* with other plastid genes. (A) K_a/K_s values of 49 loci across 20 fern plastomes plotted in descending order. The values for *ycf94* are highlighted in green for both the entire ORF and the putative transmembrane domain (asterisk). (B) Sequence variation at the first, second, and third codon positions for *Deparia matK*, *ycf94*, and *rbcL*.

positions for *ycf94* and two other plastid genes in *Deparia*, we found that patterns were comparable across all three loci, which suggests that *ycf94* substitutions are constrained in a manner consistent with it encoding a functional protein.

It is probable that *ycf94* was present in the ancestor of land plants and subsequently lost in seed plants, heterosporous lycophytes, and some moss lineages. The evolutionary history of the plastome is known to be dynamic, with genes moving both within the plastome and also to the

nuclear genome (reviewed in Cullis et al., 2009); therefore, more work will need to be done to pinpoint the origin and function of *ycf94*. Nonetheless, we have shown here that there are still undiscovered aspects to this highly conserved and well-characterized organellar genome.

ACKNOWLEDGEMENTS

The authors thank B. Castandet for discussion of the RNA editing result, N. Devos at Duke Sequencing and Genomic Technologies for consulting, and two anonymous reviewers for suggestions. Funding for this research was provided in part by the National Science Foundation (DEB-1145614 to K.M.P. and L.H.). We also acknowledge support from the GoFlag community (DEB-1541506).

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

Boudreau, E., Y. Takahashi, C. Lemieux, M. Turmel, and J. D. Rochaix. 1997. The chloroplast *ycf3* and *ycf4* open reading frames of *Chlamydomonas reinhardtii* are required for the accumulation of the photosystem I complex. *The EMBO Journal* 16: 6095–6104.

Buffalo, V. 2011. Scythe-a Bayesian adapter trimmer. Website: http://github.com/vsbuffalo/scythe.

Castandet, B., A. M. Hotto, S. R. Strickler, and D. B. Stern. 2016. ChloroSeq, an optimized chloroplast RNA-seq bioinformatic pipeline, reveals remodeling of the organellar transcriptome under heat stress. *G3: Genes Genomes, Genetics* 6: 2817–2827.

Cullis, C. A., B. J. Vorster, C. Van Der Vyver, and K. J. Kunert. 2009. Transfer of genetic material between the chloroplast and nucleus: How is it related to stress in plants? *Annals of Botany* 103: 625–633.

Dierckxsens, N., P. Mardulyn, and G. Smits. 2016. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research* 45: e18.

Duffy, A. M., S. A. Kelchner, and P. G. Wolf. 2009. Conservation of selection on *matK* following an ancient loss of its flanking intron. *Gene* 438: 17–25.

Graham, S. W., V. K. Y. Lam, and V. S. F. T. Merckx. 2017. Plastomes on the edge: The evolutionary breakdown of mycoheterotroph plastid genomes. *New Phytologist* 214: 48–55.

Gray, M. W., and P. S. Covello. 1993. RNA editing in plant mitochondria and chloroplasts. *The FASEB Journal* 7: 64–71.

Hallick, R. B., and A. Bairoch. 1994. Proposals for the naming of chloroplast genes. III. Nomenclature for open reading frames encoded in chloroplast genomes. *Plant Molecular Biology Reporter* 12: S29–30.

Hsu, P. Y., and P. N. Benfey. 2017. Small but mighty: Functional peptides encoded by small ORFs in plants. *Proteomics* 1700038.

Jansen, R. K., and T. A. Ruhlman. 2012. Plastid genomes of seed plants. *In* Genomics of Chloroplasts and Mitochondria, Advances in Photosynthesis and Respiration, 103–126. Springer, Dordrecht, The Netherlands.

Joshi, N. A., and J. N. Fass. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. Website: http://github.com/najoshi/sickle.

Kearse, M., R. Moir, A. Wilson, and S. Stones-Havas. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649.

Kikuchi, S., J. Bedard, M. Hirano, Y. Hirabayashi, M. Oishi, M. Imai, M. Takase, et al. 2013. Uncovering the protein translocon at the chloroplast inner envelope membrane. *Science* 339: 571–574.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. 2013. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* 14: R36.

Krogh, A., B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology* 305: 567–580.

Kugita, M., A. Kaneko, Y. Yamamoto, Y. Takeya, T. Matsumoto, and K. Yoshinaga. 2003. The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: Insight into the earliest land plants. *Nucleic Acids Research* 31: 716–721.

Kuo, L.-Y., A. Ebihara, W. Shinohara, G. Rouhan, K. R. Wood, C.-N. Wang, and W.-L. Chiou. 2016. Historical biogeography of the fern genus *Deparia* (Athyriaceae) and its relation with polyploidy. *Molecular Phylogenetics and Evolution* 104: 123–134.

Kuo, L.-Y., F.-W. Li, W.-L. Chiou, and C.-N. Wang. 2011. First insights into fern *matK* phylogeny. *Molecular Phylogenetics and Evolution* 59: 556–566.

Li, F.-W., L.-Y. Kuo, K. M. Pryer, and C. J. Rothfels. 2016. Genes translocated into the plastid inverted repeat show decelerated substitution rates and elevated GC content. *Genome Biology and Evolution* 8: 2452–2458.

Ohyama, K., H. Fukuzawa, T. Kohchi, H. Shirai, T. Sano, S. Sano, K. Umesono, et al. 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature* 322: 572–574.

PPG I. 2016. A community-derived classification for extant lycophytes and ferns. *Journal of Systematics and Evolution* 54: 563–603.

Rothfels, C. J., A. Larsson, F.-W. Li, E. M. Sigel, L. Huiet, D. O. Burge, M. Ruhsam, et al. 2013. Transcriptome-mining for single-copy nuclear markers in ferns. *PLoS ONE* 8: e76957.

Schallenberg-Rüdinger, M., and V. Knoop. 2016. Chapter two-coevolution of organelle RNA editing and nuclear specificity eactors in early land plants. *Advances in Botanical Research* 78: 37–93.

Shimada, H., and M. Sugiura. 1991. Fine structural features of the chloroplast genome: Comparison of the sequenced chloroplast genomes. *Nucleic Acids Research* 19: 983–995.

Shinozaki, K., M. Ohme, M. Tanaka, T. Wakasugi, N. Hayshida, T. Matsubayasha, N. Zaita, et al. 1986. The complete nucleotide sequence of the tobacco chloroplast genome. *Plant Molecular Biology Reporter* 4: 111–148.

Soltis, D. E., P. S. Soltis, and J. J. Doyle. 1998. Molecular Systematics of Plants II: DNA Sequencing. Kluwer Academic, Boston, Massachusetts, USA.

Tamura, K., G. Stecher, D. Peterson, A. Filipski, and S. Kumar. 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* 30: 2725–2729.

Wickett, N. J., S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences USA* 111: E4859–E4868.

Wolf, P. G. 2003. Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. *DNA Research* 10: 59–65.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586–1591.